09/288,256

日本国特許庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されてる事項と同一であることを証明する。
This is to certify that the annexed is a true copy of the following application as filed th this Office.

出願年月日
Date of Application: １９９８年　４月１０日

出願番号
Application Number: 平成１０年特許願第１１４４１４号

願人
Applicant (s): 株式会社リコー

CERTIFIED COPY OF
PRIORITY DOCUMENT

１９９９年　４月２３日

特許庁長官
Commissioner,
Patent Office 伊佐山建志

【書類名】　特許願

【整理番号】　9801993

【提出日】　平成10年 4月10日

【あて先】　特許庁長官殿

【国際特許分類】　G06F 17/21

【発明の名称】　文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【請求項の数】　7

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　長束　哲郎

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　宮地　達生

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　嶋田　敦夫

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　武谷　一寿

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　剣持　栄治

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　中島　明子

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　山崎　真湖人

【発明者】

【住所又は居所】　東京都大田区中馬込１丁目３番６号　株式会社リコー内

【氏名】　藤田　克彦

【特許出願人】

【識別番号】　000006747

【氏名又は名称】　株式会社リコー

【代表者】　桜井　正光

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【書類名】　明細書

【発明の名称】　文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体

【特許請求の範囲】

【請求項１】　文書の内容に基づいて文書の分類をおこなう文書分類装置において、

一つまたは複数の項目から構成された文書データを入力する入力手段と、

前記入力手段により入力された文書データを構成する前記項目を指定する指定手段と、

前記指定手段により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換手段と、

前記変換手段により変換された変換データをもちいて文書を分類する分類手段と、

を備えたことを特徴とする文書分類装置。

【請求項２】　文書の内容に基づいて文書の分類をおこなう文書分類装置において、

一つまたは複数の項目から構成された文書データを入力する入力手段と、

前記入力手段により入力された文書データを構成する前記項目を指定する指定手段と、

前記指定手段により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換手段と、

前記変換手段により変換された変換データをもちいて各文書の特徴ベクトルを生成する文書ベクトル生成手段と、

前記文書ベクトル生成手段により生成された各文書の特徴ベクトルをもちいて文書を分類する分類手段と、

を備えたことを特徴とする文書分類装置。

【請求項３】　前記変換手段は、前記文書データを変換する際、前記各項目のデータが分離可能となるように前記項目のデータ間に所定の記号を挿入することを特徴とする請求項１または２に記載の文書分類装置。

【請求項４】　文書の内容に基づいて文書の分類をおこなう文書分類方法において、

一つまたは複数の項目から構成された文書データを入力する入力工程と、

前記入力工程により入力された文書データを構成する前記項目を指定する指定工程と、

前記指定工程により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換工程と、

前記変換工程により変換された変換データをもちいて文書を分類する分類工程と、

を含んだことを特徴とする文書分類方法。

【請求項５】　文書の内容に基づいて文書の分類をおこなう文書分類方法において、

一つまたは複数の項目から構成された文書データを入力する入力工程と、

前記入力工程により入力された文書データを構成する前記項目を指定する指定工程と、

前記指定工程により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換工程と、

前記変換工程により変換された変換データをもちいて各文書の特徴ベクトルを生成する文書ベクトル生成工程と、

前記文書ベクトル生成工程により生成された各文書の特徴ベクトルをもちいて文書を分類する分類工程と、

を含んだことを特徴とする文書分類方法。

【請求項６】　前記変換工程は、前記文書データを変換する際、前記各項目のデータが分離可能となるように前記項目のデータ間に所定の記号を挿入することを特徴とする請求項４または５に記載の文書分類方法。

【請求項７】　前記請求項４～６のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

この発明は、文書の内容に基づいて文書を分類する文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体に関する。

【０００２】

【従来の技術】

従来、文書分類装置として、たとえば、特開平７－３６８９７号公報記載の文書分類装置には、文書を単語を特徴とする文書ベクトルとみなし、クラスタリング手法を用いてこれらの文書ベクトルを群分けし、文書の自動分類をおこなうものが記載されている。

【０００３】

また、通常、文書データは一般的にデータベース化されており、文書内容だけでなく作成日や作成者などの書誌的項目が付加されていたり、また文書内容自体が複数の項目を含んでいる場合が多い。たとえば、特許公報は、「特許請求の範囲」「発明の詳細な説明」といった複数の項目から構成されている。

【０００４】

【発明が解決しようとする課題】

しかしながら、上記従来技術の文書分類装置は、複数の項目を持つ文書データに対して、操作者が分類対象とする項目を任意に指定することができないことから、分類に悪影響を与えるデータが付加されていたり、また、複数の項目を組み合わせることが出来ないことから、分類に有効なデータが不足したりして、精度の高い分類結果を得ることができないという問題があった。

【０００５】

この発明は、上述した従来例による問題点を解消するため、操作者の意図が反映された精度の高い分類をおこなうことができる文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体を提供することを目的とする。

【０００６】

【課題を解決するための手段】

上述した課題を解決し、目的を達成するため、請求項１の発明に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、一つまたは複数の項目から構成された文書データを入力する入力手段と、前記入力手段により入力された文書データを構成する前記項目を指定する指定手段と、前記指定手段により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換手段と、前記変換手段により変換された変換データをもちいて文書を分類する分類手段と、を備えたことを特徴とする。

【０００７】

この請求項１の発明によれば、文書を分類する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が期待する分類の観点に必要と思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類を効率よくおこなうことが可能である。

【０００８】

また、請求項２に係る文書分類装置は、文書の内容に基づいて文書の分類をおこなう文書分類装置において、一つまたは複数の項目から構成された文書データを入力する入力手段と、前記入力手段により入力された文書データを構成する前記項目を指定する指定手段と、前記指定手段により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換手段と、前記変換手段により変換された変換データをもちいて各文書の特徴ベクトルを生成する文書ベクトル生成手段と、前記文書ベクトル生成手段により生成された各文書の特徴ベクトルをもちいて文書を分類する分類手段と、を備えたことを特徴とする。

【０００９】

この請求項２の発明によれば、文書を分類するための各文書の特徴ベクトルを生成する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が期待する分類の観点に必要と思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能である。

【００１０】

また、請求項３に係る文書分類装置は、請求項１または２の発明において、前記変換手段が、前記文書データを変換する際、前記各項目のデータが分離可能となるように前記項目のデータ間に所定の記号を挿入することを特徴とする。

【００１１】

この請求項３の発明によれば、各変換データの間に区切りとなる記号を挿入するので、形態素解析等の解析処理をおこなう際に、各項目に対応するデータをそのまま結合させることにより変換データ全体として全く別の意味が構成されることを回避することが可能である。

【００１２】

また、請求項４に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、一つまたは複数の項目から構成された文書データを入力する入力工程と、前記入力工程により入力された文書データを構成する前記項目を指定する指定工程と、前記指定工程により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換工程と、前記変換工程により変換された変換データをもちいて文書を分類する分類工程とを含んだことを特徴とする。

【００１３】

この請求項４の発明によれば、文書を分類する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が自分が期待する分類の観点に必要と思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能である。

【００１４】

また、請求項５に係る文書分類方法は、文書の内容に基づいて文書の分類をおこなう文書分類方法において、一つまたは複数の項目から構成された文書データを入力する入力工程と、前記入力工程により入力された文書データを構成する前記項目を指定する指定工程と、前記指定工程により指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換工程と、前記変換工

程により変換された変換データをもちいて各文書の特徴ベクトルを生成する文書ベクトル生成工程と、前記文書ベクトル生成工程により生成された各文書の特徴ベクトルをもちいて文書を分類する分類工程と、を含んだことを特徴とする。

【００１５】

この請求項５の発明によれば、文書を分類するための各文書の特徴ベクトルを生成する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が自分が期待する分類の観点に必要と思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能である。

【００１６】

また、請求項６に係る文書分類方法は、請求項４または５の発明において、前記変換工程が、前記文書データを変換する際、前記各項目のデータが分離可能となるように前記項目のデータ間に所定の記号を挿入することを特徴とする。

【００１７】

この請求項６の発明によれば、各変換データの間に区切りとなる記号を挿入するので、形態素解析等の解析処理の際に、複数の項目のデータを一つのデータとして混同して扱われることを回避できるとともに、各項目ごとの内容データが瞬時に識別することが可能である。

【００１８】

また、請求項７の発明に係る記憶媒体は、請求項４～６に記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項４～６の動作をコンピュータによって実現することが可能である。

【００１９】

【発明の実施の形態】

以下に添付図面を参照して、この発明に係る文書分類装置、文書分類方法およびその方法をコンピュータに実行させるプログラムを記録したコンピュータ読み取り可能な記録媒体の好適な実施の形態を詳細に説明する。

【００２０】

（実施の形態１）

まず、この発明の実施の形態１による文書分類装置を構成する情報処理システム全体のハードウエア構成を説明する。図１は、実施の形態１による文書分類装置を構成する情報処理システム全体のハードウエア構成を示す説明図である。図１において、実施の形態１による文書分類装置を構成する情報処理システムは、サーバー／クライアント方式で構成されている。すなわち、サーバー１０１と複数のクライアント１０２がネットワーク１０３によって接続されている。

【００２１】

クライアント１０２は、分類データの生成、サーバー１０１への指示、分類結果の表示などをおこなう。一方、クライアント１０２からの指示に従い、サーバー１０１は文書（テキスト）分類に関する処理を膨大な数値演算によりおこない、その処理の結果をクライアント１０２へ送る。より具体的には、サーバー１０１においては、テキスト分類処理がおこなわれ、クライアント１０２においては、分類データ生成、処理実行指示、テキスト分類結果表示等がおこなわれる。

【００２２】

また、サーバー１０１とクライアント１０２との間のデータのやりとりはファイル共有という方法をもちいる。すなわち、分類処理にもちいるファイルをサーバー１０１上の共有フォルダに作成することにより両者はデータのやりとりをおこなう。したがって、クライアント１０２からはサーバー１０１の共有フォルダをネットワーク共有して利用することが可能である。

【００２３】

つぎに、サーバー１０１およびクライアント１０２のハードウエア構成について説明する。図２は、実施の形態１による文書分類装置を構成する情報処理システムにおけるサーバー１０１をハードウエア的に示す説明図である。サーバー１０１は、たとえばワークステーション（ＷＳ）等がもちいられる。

【００２４】

図２において、２０１はサーバー１０１全体を制御するＣＰＵを、２０２はブートプログラム等を記憶したＲＯＭを、２０３はＣＰＵ２０１のワークエリアとして使用されるＲＡＭ２０３を、２０４は通信回線２０５を介してネットワーク

１０３に接続され、そのネットワーク１０３と内部のインターフェイスを司るインターフェイス（Ｉ／Ｆ）を、２０６はデータを記憶するディスク装置を示している。２００は上記各部を結合させるためのバスを示している。

【００２５】

そのほか、文書情報、画像情報、機能情報等を表示するディスプレイ２０８や、データを入力するためのキーボード２０９およびマウス２１０等が同様に接続されていてもよい。さらに、ディスク装置２０６には、クライアント１０２との間のデータのやりとりをするための共有フォルダ２０７が設けられている。

【００２６】

また、図３は、実施の形態１による文書分類装置を構成する情報処理システムにおけるクライアント１０２をハードウエア的に示す説明図である。クライアント１０２は、たとえばパーソナルコンピュータ（ＰＣ）等がもちいられる。

【００２７】

図３において、３０１はシステム全体を制御するＣＰＵを、３０２はブートプログラム等を記憶したＲＯＭを、３０３はＣＰＵ３０１のワークエリアとして使用されるＲＡＭを、３０４はＣＰＵ３０１の制御にしたがってＨＤ（ハードディスク）３０５に対するデータのリード／ライトを制御するＨＤＤ（ハードディスクドライブ）を、３０５はＨＤＤ３０４の制御で書き込まれたデータを記憶するＨＤを、３０６はＣＰＵ３０１の制御にしたがってＦＤ（フロッピーディスク）３０７に対するデータのリード／ライトを制御するＦＤＤ（フロッピーディスクドライブ）を、３０７はＦＤＤ３０６の制御で書き込まれたデータを記憶する着脱自在のＦＤを、３０８はドキュメント、画像、機能情報等を表示するディスプレイをそれぞれ示している。

【００２８】

また、３０９は通信回線３１０を介してネットワーク１０３に接続され、そのネットワーク１０３と内部のインターフェイスを司るインターフェイス（Ｉ／Ｆ）を、３１１は文字、数値、各種指示等の入力のためのキーを備えたキーボードを、３１２はカーソルの移動や範囲選択、あるいは表示画面に表示されたアイコンやボタンの押下やウインドウの移動やサイズの変更等をおこなうマウスを、３

１３はＯＣＲ（Ｏｐｔｉｃａｌ　Ｃｈａｒａｃｔｅｒ　Ｒｅａｄｅｒ）機能を備えた画像を光学的に読み取るスキャナを、３１４は分類結果を含むデータの内容等を印刷するプリンタを、３１５は上記各部を結合するためのバスをそれぞれ示している。

【００２９】

つぎに、実施の形態１による文書分類装置の機能的構成について説明する。図４は実施の形態１による文書分類装置の構成を機能的に示すブロック図である。図４において、文書分類装置は、入力部４０１と、指定部４０２と、変換部４０３と、変換データ記憶部４０４と、分類部４０５と、分類結果記憶部４０６を含む構成である。

【００３０】

つぎに、各構成部についてその内容を詳細に説明する。なお、入力部４０１、指定部４０２、変換部４０３、変換データ記憶部４０４、分類部４０５、分類結果記憶部４０６は、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログラムに記載された命令に従ってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現するものである。

【００３１】

（入力部４０１）

入力部４０１は、文書データを入力するものであり、たとえば、キーボード２０９または３１１、ＯＣＲ機能を備えたスキャナ３１３、またはネットワーク１０３を経由して文書や文書群を得ることができるＩ／Ｆ２０４または３０９等である。また、入力部４０１は、上記以外に文書データを取得することができるものであれば、それらすべてを含む。

【００３２】

たとえば、文書データがデータベース化されている場合に、そのデータベースが記録された媒体を本実施の形態の文書分類装置に組み入れた場合も文書データの入力とする。さらに、入力した文書データを記憶する図示しない文書データ記憶部を含んでいてもよい。この文書データ記憶部は、たとえば大容量のメモリを

有するサーバー１０１のディスク装置２０６等であってもよい。

【００３３】

ここで、文書とは、本実施の形態にあっては、自然言語で記述された一つ以上の文の集まりであり、それが分類対象となる場合はこれを文書という。具体的には、公開特許公報や特定の新聞記事も文書であり、また、請求項や特定の一文を取り出したものであっても、これを文書と見なすものである。

。

【００３４】

（指定部４０２）

指定部４０２は、文書データの項目を指定するものである。指定部は、具体的には３つの処理から構成される。

【００３５】

まず、入力部４０１により入力された文書データから項目を抽出する（第１処理）。項目を抽出する方法としては、あらかじめ所定の符号（たとえば、「［」「］」等）が付されている項目を検索し、その項目を選択する等の方法がある。

【００３６】

上記第１処理は、指定部４０２でおこなう代わりに、入力部４０１においておこなってもよい。すなわち、入力部４０１が文書データを入力する際に、あわせてその文書データの項目の抽出をおこなう。その抽出結果は文書データと対応付けされて上記文書データ記憶部に記憶される。この場合は、当該抽出結果をもちいることにより、指定部４０２においては上記第１処理は省略されることになる。また、データベースの種類によってはあらかじめ項目に関する情報を有しているものがあり、その項目に関する情報を利用することによっても、上記第１処理は省略される。

【００３７】

つぎに、第１処理による項目の抽出結果、上記文書データ記憶部に記憶された上記抽出結果、または上記項目に関する情報等に基づいて、抽出された各項目がどのような項目であり、その項目に対応する内容はどのようなものであるかの一覧を操作者に提示する（第２処理）。提示の方法としては、ディスプレイ２０８

または３０８に項目のみを、あるいは項目とその項目に対応する内容の全部または一部を表示する方法等がある。

【００３８】

項目のみを表示する方法としては、たとえば、項目名を文書データ中の出現順序に基づいて横書で縦一列になるように羅列して表示するといった方法がある。この場合、表示画面上の表示行数よりも項目数が多くなる場合は、折り返して縦二列以上で表示してもよく、また、縦一列で表示して、表示画面を縦方向にスクロールできるようにしてもよい。

【００３９】

項目とその項目に対応する内容の全部または一部を表示する方法としては、たとえば、上述の項目のみを表する方法と同様に、項目名を文書データ中の出現順序に基づいて横書で縦一列になるように羅列して表示し、さらにその右側に項目名と対応して配置される位置に同じく横書でその内容を表示するといった方法がある。この場合、表示画面上の表示列数よりも内容のデータ量が多くなる場合は、表示画面を横方向にスクロールできるようにしてもよい。

【００４０】

また、項目とその項目に対応する内容の全部または一部を表示する別の方法としては、項目名のみを表示し、項目名が表示されている領域にカーソルを移動させ、所定の操作（マウス２１０または３１２のボタンあるいはキーボード２０９または３１１等の所定キーの押下）により、内容のデータの全部または一部をポップアップして表示するようにしてもよい。

【００４１】

つぎに、操作者の指示に従って、提示（表示）された項目の中から分類処理の対象となる項目を一つまたは二つ以上を同時に指定する（第３処理）。指定の方法としては、キーボード２０９または３１１やマウス２１０または３１２等のポインティングデバイスからの指定に関する指示信号に基づいて、提示されている項目の中から該当する項目を指定する。

【００４２】

この際、項目の指定は一つであってもよく、また、二つ以上を同時に指定して

もよい。また、結合の形態を併せて指定することもできる。さらに、指定の順序により、データの変換後の内容データの配列順を指定するようにしてもよい。

【００４３】

（変換部４０３）

変換部４０３は、入力された文書データを前記指定部４０２により指定された項目に対応するデータのみの内容となるように文書データを変換するものである。具体的には、文書データ中の指定された項目に対応するデータだけを抽出し、抽出されたデータのみからなる変換データへ変換するものである。

【００４４】

変換データは、単にもとの文書データにおける指定された項目の順序で各項目に対応するデータを羅列することにより変換される場合のみならず、たとえば指定された項目のデータ内容を文字列として結合して指定された項目のデータ内容だけを含む変換データとすることや、項目の順序をもとの文書データ内における順序と異なる順序に入れ替えてからデータを結合するように変換してもよい。

【００４５】

また、変換部４０３は、変換データにおける各項目のデータが分離可能となるように項目のデータ間に所定の分離記号６０１を挿入する。これにより、各項目に対応するデータの切れ目を瞬時に把握することができる。

【００４６】

また、この分離記号６０１は、形態素解析等の自然言語解析をおこなう場合に特に重要である。各項目に対応するデータが文の体をなしている場合（文の終わりが句点で終わっている場合）は、この分離記号がなくても文と文の切れ目を判断することができるが、各項目に対応するデータが文の体をなしていない場合（箇条書きの文、文の途中で項目が変わる等の場合）は、そのままデータ同士を結合させると、項目によっては、全く別の意味が構成されてしまう場合がある。そのような場合を回避するためにこの分離記号６０１を挿入する。

【００４７】

分離記号６０１は、一般的には、切れ目を表す「／（スラッシュ）」がもちいられるが、変換データ中に「／」が存在する場合には、データの「／」との混同

が生じるので、別の記号をもちいることができる。また、この記号を挿入するか否かについてキーボード２０９または３１１に所定のキーを割付け、そのキーを押下するごとに、あるいは表示画面上に所定のアイコンを表示させて、マウス２１０または３１２によりそのアイコンをクリックするごとに、分離記号６０１を挿入する／挿入しないを交互に設定するようにしてもよい。

【００４８】

（変換データ記憶部４０４）

変換データ記憶部４０４は、変換データを記憶する記憶部である。変換データ記憶部４０４としては、たとえば、サーバー１０１のディスク装置２０６またはクライアント側のハードディスク３０５、またはフロッピーディスク３０７等、変換データの容量の違いあるいは用途の違いにより、それぞれ設定することが可能である。

【００４９】

変換データ記憶部４０４には、項目の設定順序等を含む変換データのほか、前記分離記号６０１等も記憶される。変換データ記憶部４０４に記憶された変換データは、別の分類の際に用いる等、活用を図ることができる。

【００５０】

（分類部４０５）

分類部４０５は、変換部４０３により変換された変換データまたは変換データ記憶部４０４に記憶されている変換データの内容にしたがって自動的に分類する。分類部４０５については、たとえば特開平７－３６８９７号公報に開示された「文書分類装置」など従来の文書分類方法を用いて文書を分類することができる。

【００５１】

（分類結果記憶部４０６）

分類結果記憶部４０６は、分類部４０５により分類された結果を記憶する記憶部である。分類結果記憶部４０６としては、変換データ記憶部４０４と同様に、たとえば、サーバー１０１のディスク装置２０６またはクライアント側のハードディスク３０５、またはフロッピーディスク３０７等、変換データの容量の違い

あるいは用途の違いにより、それぞれ設定することが可能である。

【００５２】

つぎに、文書データと文書データを変換した変換データの一例について説明する。図５は文書データとその変換データの一例を示す説明図である。図５において、文書群として特許公報群をもちいた場合であり、５０１は文書データの一例であり、５０２は変換データの一例である。

【００５３】

文書データ５０１は、「出願番号」、「出願日」、「発明者」、「発明の名称」、「目的」、「構成」、「請求項１」、「従来技術」、「課題を解決するための手段」、「作用」、「実施例」、「発明の効果」等の項目が含まれている。

【００５４】

従来の文書分類装置では各文書データをひとまとまりとして取り扱うので、複数の項目を含む文書データに対してはすべての項目の内容データが分類処理の対象となり、操作者が望む分類の観点に不必要、あるいは悪影響を与える項目も含まれる場合がある。

【００５５】

本実施の形態においては、分類をおこなう操作者は自分が望む分類の観点に必要と思われる項目を１つ以上指定することができる。たとえば特許公報文書群の分類をおこなう際に、操作者が「発明の課題」に注目したい場合は、「目的」、「課題を解決するための手段」、「作用」、「発明の効果」を指定する。また、解決手段に注目したい場合は、「課題を解決するための手段」および「実施例」を指定することができる。分類の対象となる項目が指定されると、指定された項目に基づいて文書データを変換する。

【００５６】

図５にあっては、操作者が「目的」、「課題を解決するための手段」、「作用」、「発明の効果」の項目を指定した場合において、指定された項目の内容データだけを含むように変換した場合の例である。

【００５７】

変更データ５０２から明らかなように、「目的」の項目に対応するデータであ

る「履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。」と、「課題を解決するための手段」の項目に対応するデータである「上記目的を達成するために・・・表示する表示手段とを有する。」と、「作用」の項目に対応するデータである「以上の構成において、入力手続きより・・・表示するように動作する。」と、「発明の効果」の項目に対応するデータである「以上説明したよう日本発明によれば・・・再現できる効果がある。」とが結合して一つの文書を構成している。

【００５８】

また、図６は、同一の文書データをもちいて、操作者が「目的」、「課題を解決するための手段」、「作用」、「発明の効果」を指定した場合において、指定された項目の内容データだけを含み、各項目のデータ間に分離記号６０１（「／」）を挿入するように変換した場合の例である。

【００５９】

つぎに、実施の形態１による文書分類装置の一連の処理の手順について説明する。図７は実施の形態１による文書分類装置の一連の処理の手順を示すフローチャートである。

【００６０】

図７のフローチャートにおいて、まず、入力部１は文書データの入力をおこなう（ステップＳ７１０）。また、指定部４０２は項目の指定をおこなう（ステップＳ７２０）。

【００６１】

変換部４０３は、ステップＳ７１０において入力された文書データをステップＳ７２０において指定されて項目の内容になるように変換データへ変換する（ステップＳ７３０）。また、必要に応じて分離記号６０１を項目に対応するデータ間に挿入する（ステップＳ７４０）。変換された変換データは、分離記号データとともに変換データ記憶部４０４により記憶される（ステップＳ７５０）。

【００６２】

上記ステップＳ７３０において変換された変換データあるいは上記ステップＳ７５０において変換データ記憶部４０４によって記憶された変換データに基づい

て、分類部４０５は文書の分類をおこなう（ステップＳ７８０）。分類処理が終了後、分類処理の結果は分類結果記憶部４０６により記憶され（ステップＳ７９０）、すべての処理は終了する。

【００６３】

以上説明したように、実施の形態１によれば、指定された項目により文書データが変換データへ変換され、その変換データに基づいて文書の分類をおこなうので、その他の不要な項目の内容による分類結果への影響を抑制することができる。また、分離記号６０１の挿入により、変換データにおける結合された項目ごとのデータの識別ができ、かつ、項目間のデータの結合による内容の混同を回避することができる。

【００６４】

（実施の形態２）

さて、実施の形態１では、変換データをもちいて文書を分類したが、以下に説明する実施の形態２のように、変換データを用いて文書の特徴ベクトルを生成し、その特徴ベクトルを用いて文書を分類するようにしてもよい。

【００６５】

まず、実施の形態２による文書分類装置の機能的構成について説明する。図８は、実施の形態２による文書分類装置の構成を機能的に示すブロック図である。図８において、実施の形態１の図４と同一のものに関しては同じ番号を付して、その説明を省略する。

【００６６】

図８において、文書分類装置は、入力部４０１と、指定部４０２と、変換部４０３と、変換データ記憶部４０４と、分類部４０５と、分類結果記憶部４０６のほかに、文書ベクトル生成部８０１と、文書ベクトル記憶部８０２とを含む構成である。

【００６７】

なお、文書ベクトル生成部８０１と文書ベクトル記憶部８０２は、他の構成部と同様に、ＲＯＭ２０２または３０２、ＲＡＭ２０３または３０３、あるいはディスク装置３０６またはハードディスク３１６等の記録媒体に記録されたプログ

ラムに記載された命令に従ってＣＰＵ２０１または３０１等が命令処理を実行することにより、各部の機能を実現するものである。

【００６８】

（文書ベクトル生成部８０１）

文書ベクトル生成部８０１は、各文書の特徴ベクトルを生成する。文書の特徴ベクトルを生成するためには、文書データに対して形態素解析等の自然言語解析処理をおこなう必要がある。この自然言語解析処理は、図示しない文書解析部によって、各文書データについて各項目ごとおこなわれる。形態素解析は従来の形態素解析手法を用いることができる。

【００６９】

文書ベクトル生成部８０１では各文書データに対して前記文書解析部によって得られた解析結果を用いて文書ベクトルを生成するものである。この際に指定部４０２によって指定された項目に関する解析結果のみに基づいて文書ベクトルの生成をおこなう。たとえば各文書データに対して指定部４０２で指定された項目の内容データから得られる特徴ベクトルだけを加算して文書ベクトルを生成することで、指定部４０２で指定された項目の内容データだけを反映した文書ベクトルを生成することができる。

【００７０】

（文書ベクトル記憶部８０２）

文書ベクトル記憶部８０２は、文書ベクトル生成部８０１によって生成された各文書の特徴ベクトルを記憶する記憶部である。文書ベクトル記憶部８０２においては同一文書であっても指定部４０２により指定される項目によっては、その文書の特徴ベクトルが異なってくるので、指定ごとにそれぞれ文書の特徴ベクトルを記憶する。分類部４０５による文書の分類をおこなう際には、あらかじめ文書ベクトル記憶部８０２によって記憶された上記文書の特徴ベクトルをもちいるので、効率よく文書の分類をおこなうことができる。

【００７１】

文書ベクトル記憶部８０２としては、たとえば、サーバー１０１のディスク装置２０６またはクライアント側のハードディスク３０５、またはフロッピーディ

スク３０７等を、変換データの容量の違いあるいは用途の違いにより、それぞれ設定することが可能である。

【００７２】

（分類部４０５）

分類部４０５は、変換部４０３により変換された各文書の特徴ベクトル間の類似度に基づいて文書を分類するものである。具体的には、生成された分類対象データに対して、カイ自乗法の手法、判別分析の手法、およびクラスタ分析の手法等の分類手法を適用することで、文書分類をおこなうことができる。ここではベクトルデータが適用できる分類手法であれば、その手法は問わない。

【００７３】

つぎに、実施の形態２による文書分類装置の一連の処理の手順について説明する。図９は実施の形態２による文書分類装置の一連の処理の手順を示すフローチャートである。図９のフローチャートにおいて、ステップＳ７１０～Ｓ７５０までは、実施の形態１の図７のフローチャートと同一ステップなので、同一ステップ番号を付して、その説明は省略する。

【００７４】

上記ステップＳ７３０において変換された変換データあるいは上記ステップＳ７５０において記憶された変換データに基づいて、文書ベクトル生成部８０１は各文書の特徴ベクトルの生成をおこなう（ステップＳ７６０）。生成された各文書の特徴ベクトルは文書ベクトル記憶部８０２により記憶される（ステップＳ７７０）。

【００７５】

上記ステップＳ７６０において変換された変換データあるいはステップＳ７７０において記憶された変換データに基づいて、分類部４０５は文書の分類がおこなう（ステップＳ７８０）。分類処理が終了後、分類処理の結果は分類結果記憶部４０６により記憶され（ステップＳ７９０）、すべての処理は終了する。

【００７６】

以上、実施の形態２によれば、指定された項目により文書データが変換データへ変換され、変換データに基づいて、各文書の特徴ベクトルの生成をおこなうの

で、操作者の意図をより反映した文書の特徴ベクトルを用いて文書の分類をおこなうことでき、その他の不要な項目の内容による分類結果への影響を抑制することができる。

【００７７】

【発明の効果】

以上説明したように、請求項１の発明によれば、文書を分類する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が期待する分類の観点に必要と思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類を効率よくおこなうことが可能な文書分類装置が得られるという効果を奏する。

【００７８】

また、請求項２の発明によれば、文書を分類するための各文書の特徴ベクトルを生成する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が期待する分類の観点に必要と思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能な文書分類装置が得られるという効果を奏する。

【００７９】

また、請求項３の発明によれば、各変換データの間に区切りとなる記号を挿入するので、形態素解析等の解析処理の際に、複数の項目のデータを一つのデータとして混同して扱われることを回避できるとともに、各項目ごとの内容データが瞬時に識別することが可能な文書分類装置が得られるという効果を奏する。

【００８０】

また、請求項４の発明によれば、文書を分類する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が自分が期待する分類の観点に必要と思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【００８１】

また、請求項５の発明によれば、文書を分類するための各文書の特徴ベクトルを生成する際に、指定された項目の内容データだけが用いられるので、その他の項目の内容による分類結果への影響を防ぐことができる。そのため、操作者が自分が期待する分類の観点に必要と思われる文書データの項目を指定することにより、操作者が望む分類により近い精度の高い分類をおこなうことが可能な文書分類方法が得られるという効果を奏する。

【００８２】

また、請求項６の発明によれば、各変換データの間に区切りとなる記号を挿入するので、形態素解析等の解析処理の際に、複数の項目のデータを一つのデータとして混同して扱われることを回避できるとともに、各項目ごとの内容データが瞬時に識別することが可能な文書分類方法が得られるという効果を奏する。

【００８３】

また、請求項７の発明によれば、請求項４～６のいずれか一つに記載された方法をコンピュータに実行させるプログラムを記録したことで、そのプログラムを機械読み取り可能となり、これによって、請求項４～６の動作をコンピュータによって実現することが可能な記録媒体が得られるという効果を奏する。

【図面の簡単な説明】

【図１】

この発明の実施の形態１による文書分類装置を構成する情報処理システム全体のハードウエア構成を示す説明図である。

【図２】

実施の形態１による文書分類装置を構成する情報処理システムにおけるサーバーをハードウエア的に示す説明図である。

【図３】

実施の形態１による文書分類装置を構成する情報処理システムにおけるクライアントをハードウエア的に示す説明図である。

【図４】

実施の形態１による文書分類装置の構成を機能的に示すブロック図である。

【図５】

実施の形態１による文書分類装置における文書データおよび変換データの内容の一例を示す説明図である。

【図６】

実施の形態１による文書分類装置における文書データおよび変換データの内容の別の一例を示す説明図である。

【図７】

実施の形態１による文書分類装置の一連の処理の手順を示すフローチャートである。

【図８】

この発明の実施の形態２による文書分類装置の構成を機能的に示すブロック図である。

【図９】

実施の形態２による文書分類装置の一連の処理の手順を示すフローチャートである。

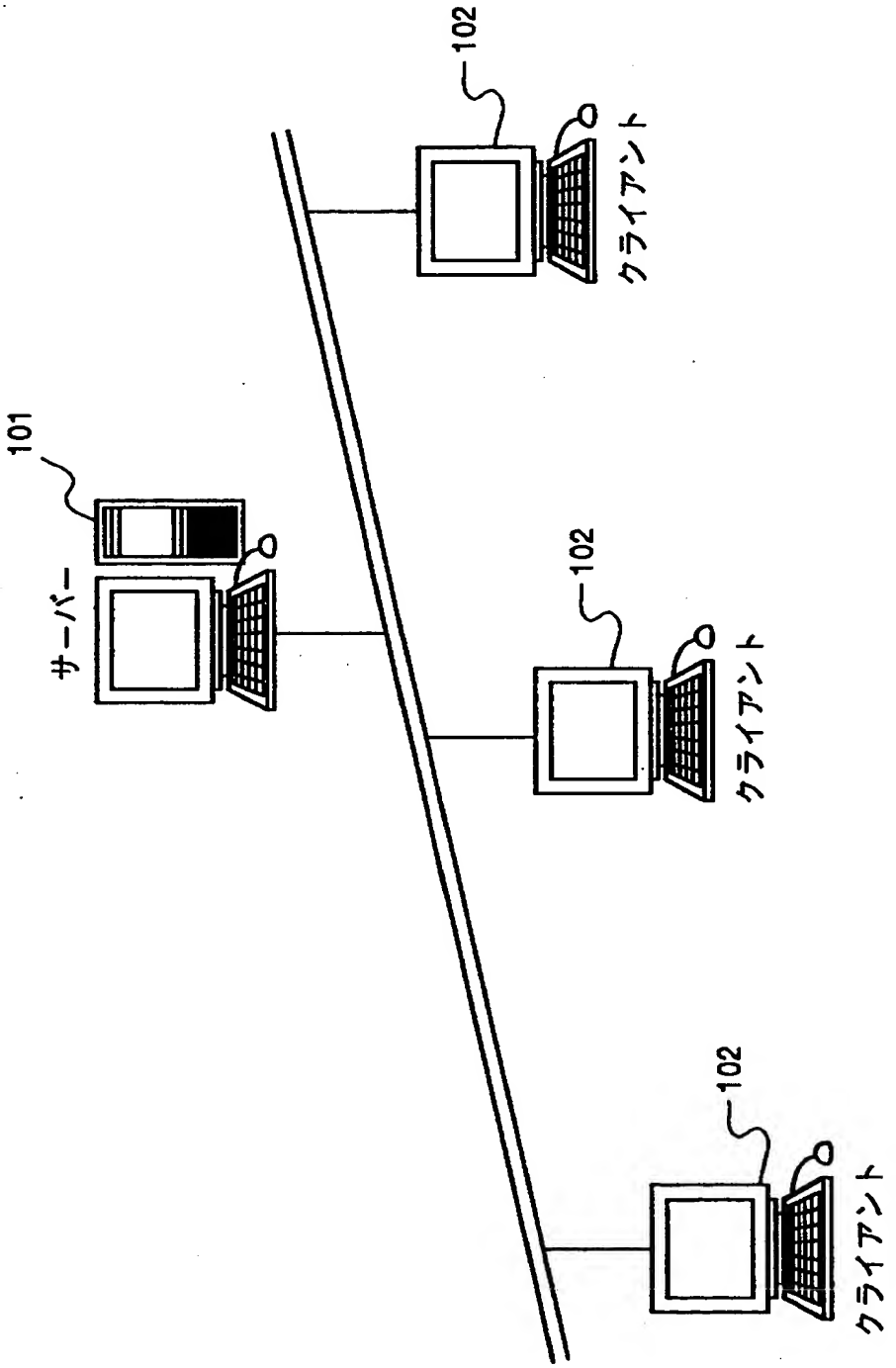
【符号の説明】

１０１　サーバー
１０２　クライアント
１０３　ネットワーク
２０１　ＣＰＵ
２０４　Ｉ／Ｆ
２０６　ディスク装置
３０１　ＣＰＵ
３０６　ハードディスク
３０８　ディスプレイ
３０９　Ｉ／Ｆ
３１１　キーボード
３１２　マウス
３１３　スキャナ

４０１　入力部
４０２　指定部
４０３　変換部
４０４　変換データ記憶部
４０５　分類部
４０６　分類結果記憶部
５０１　文書データ
５０２　変換データ
６０１　分離記号
８０１　文書ベクトル生成部
８０２　文書ベクトル記憶部

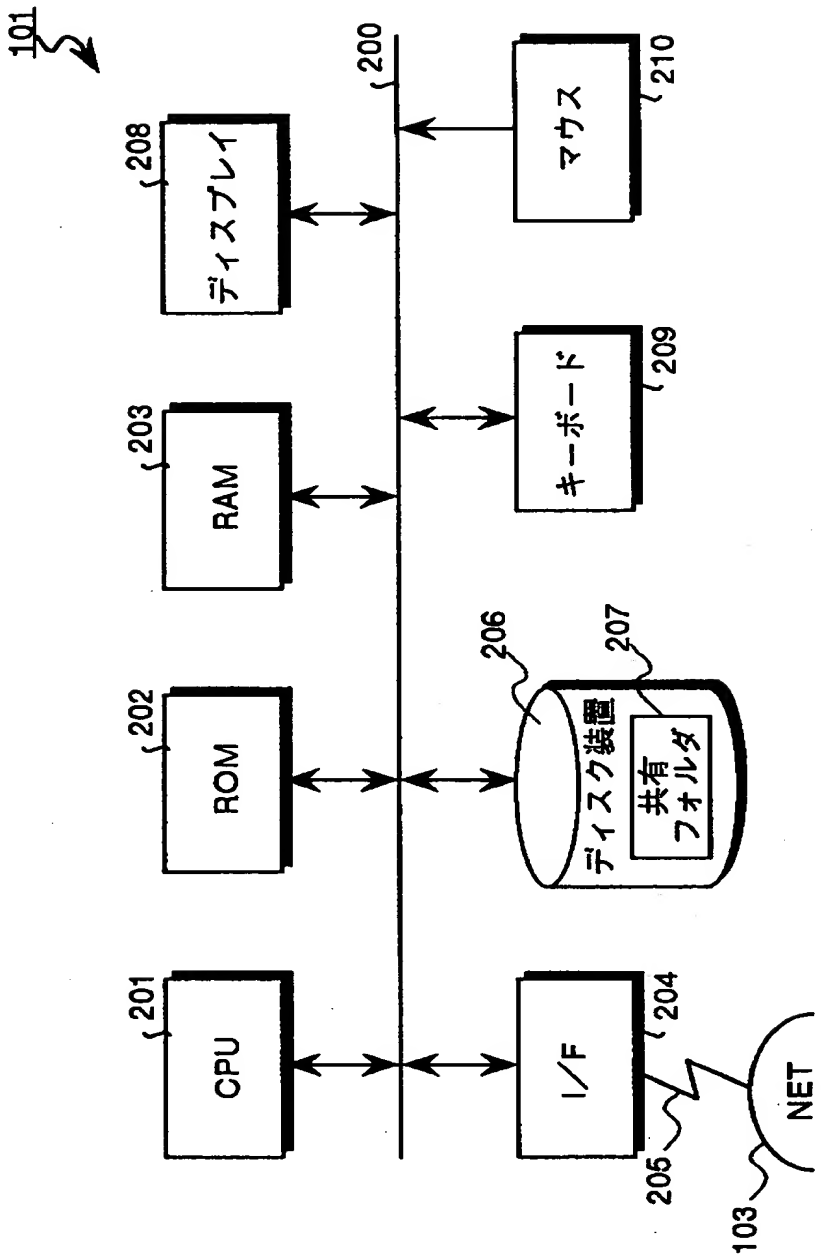
【書類名】　図面

【図１】

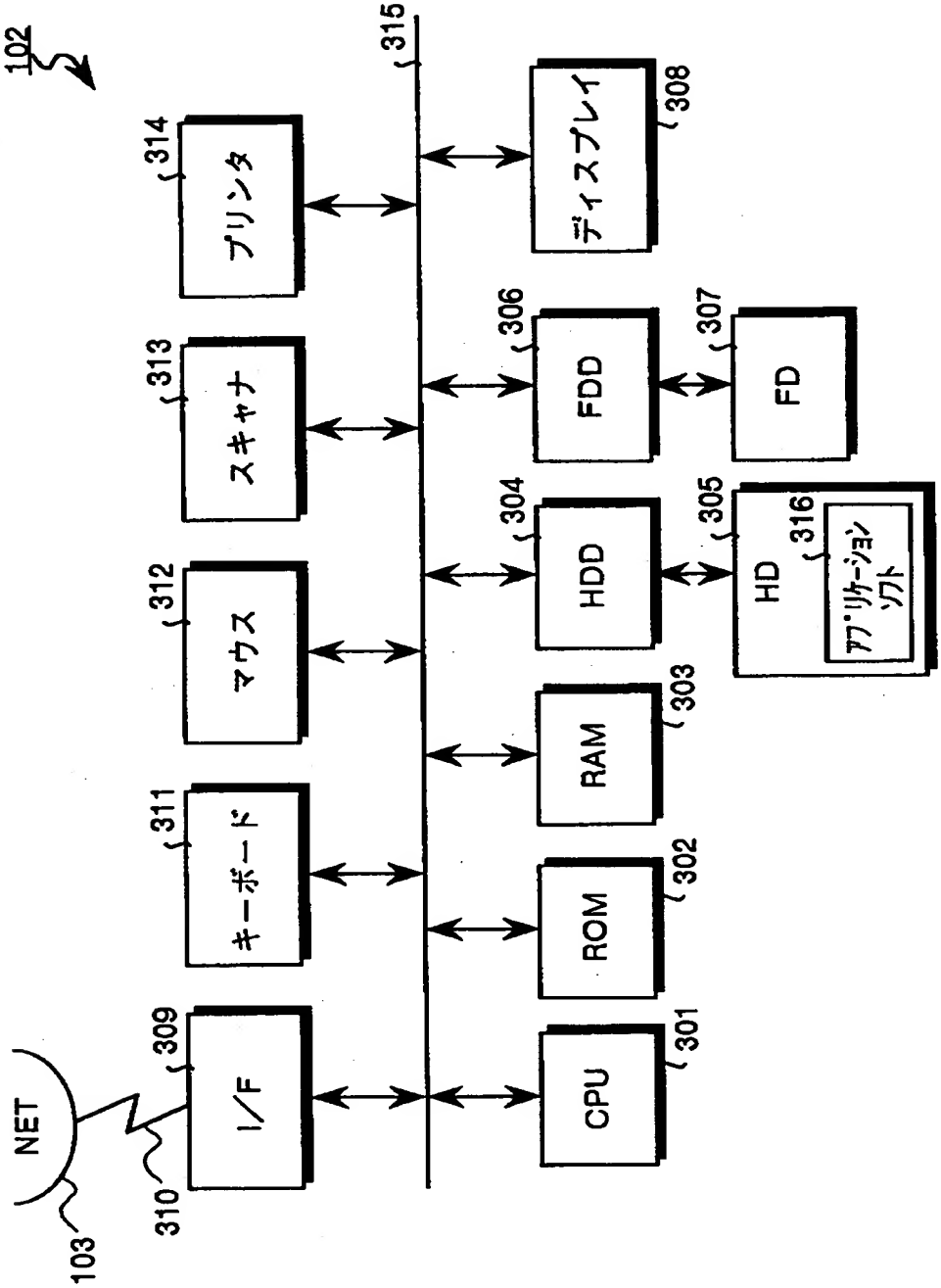


出証特平１１－３０２４８４８

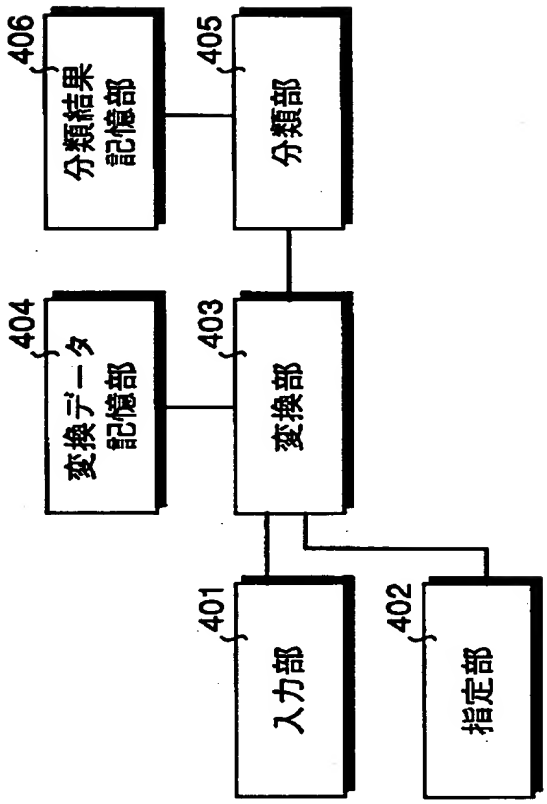
【図２】


101
200
208
ディスプレイ
マウス
210
203
RAM
キーボード
209
206
207
202
ROM
ディスク装置
共有
フォルダ
201
CPU
I/F
204
205
NET
103

【図３】


102
315
314
プリンタ
308
ディスプレイ
313
スキャナ
306
FDD
307
FD
312
マウス
304
HDD
305
HD
316
アプリケーション
ソフト
311
キーボード
303
RAM
309
I/F
302
ROM
NET
310
301
CPU
103

【図４】



【図５】

501

出願番号
特願平3-000000

出願日
平成3年（1991）○月○日

発明者
山田太郎

発明の名称
情報処理装置

目的
履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。

構成
入力部7より 入力された・・・表示部24に表示される

請求項１
マルチウィンドウを表示して・・・特徴とする情報処理装置

従来技術
図2は、一般的な・・・表示することができる。

課題を解決するための手段
上記目的を達成するために・・・表示する表示手段とを有する。

作用
以上の構成において、入力手段より・・・表示するよう動作する。

実施例
以下、添付図面を参照して・・・ことが可能になる。

発明の効果
以上説明したように本発明によれば、・・・再現できる効果がある。

・・・

502

履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。上記目的を達成するために・・・表示する表示手段とを有する。以上の構成において、入力手段より・・・表示するように動作する。以上説明したように本発明によれば・・・再現できる効果がある。

【図６】

出願番号
特願平3-000000

出願日
平成3年（1991）○月○日

発明者
山田太郎

発明の名称
情報処理装置

目的
履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。

構成
入力部7より 入力された・・・表示部24に表示される

請求項１
マルチウィンドウを表示して・・・特徴とする情報処理装置

従来技術
図2は、一般的な・・・表示することができる。

課題を解決するための手段
上記目的を達成するために・・・表示する表示手段とを有する。

作用
以上の構成において、入力手段より・・・表示するように動作する。

実施例
以下、添付図面を参照して・・・ことが可能になる。

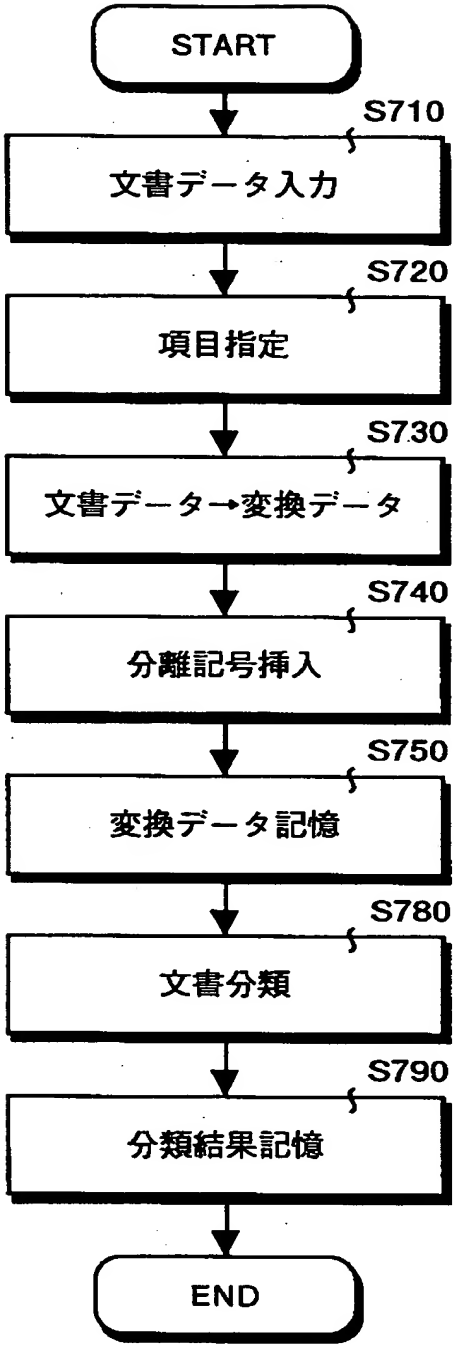
発明の効果
以上説明したように本発明によれば、・・・再現できる効果がある。

・・・

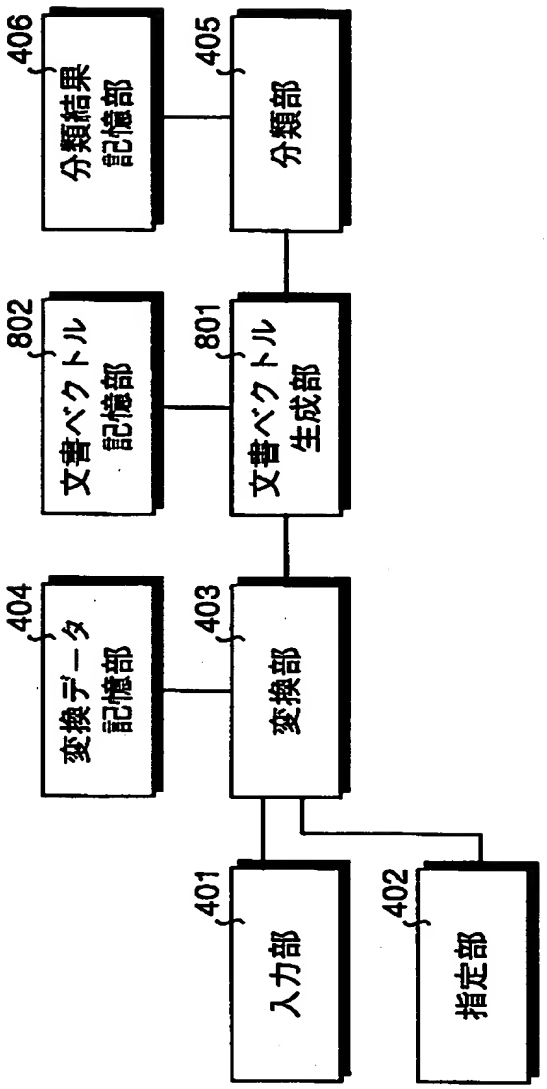
601

履歴とともに対応する画面情報を記憶しておき・・・ことを目的とする。／上記目的を達成するために・・・表示する表示手段とを有する。／以上の構成において、入力手段より・・・表示するように動作する。／以上説明したように本発明によれば・・・再現できる効果がある。

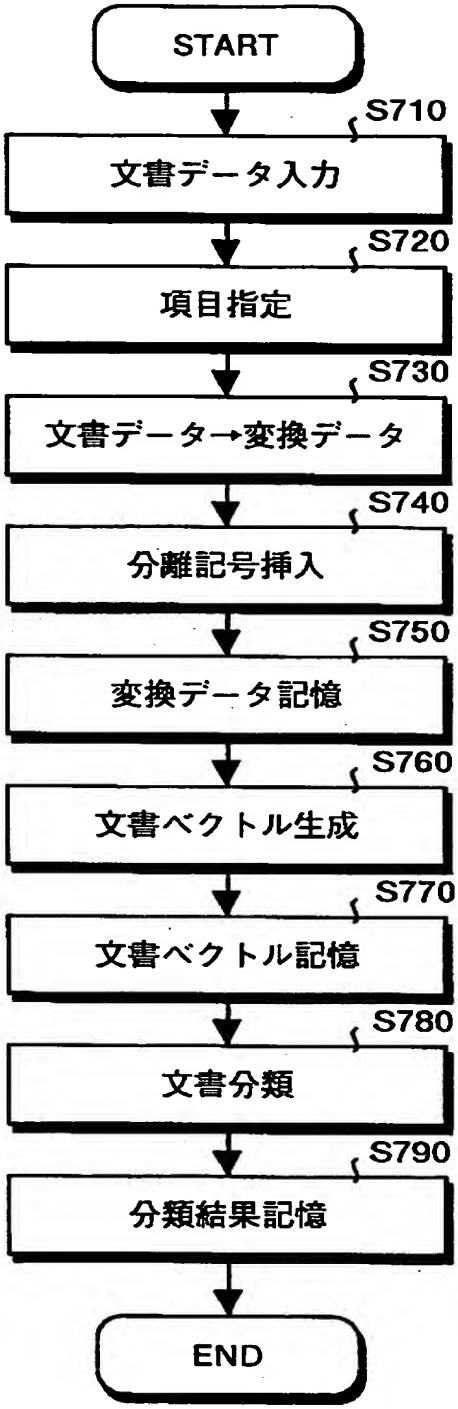
【図７】


START
S710
文書データ入力
S720
項目指定
S730
文書データ→変換データ
S740
分離記号挿入
S750
変換データ記憶
S780
文書分類
S790
分類結果記憶
END

【図８】


406
分類結果
記憶部
405
分類部
802
文書ベクトル
記憶部
801
文書ベクトル
生成部
404
変換データ
記憶部
403
変換部
401
入力部
402
指定部

【図９】


START
S710
文書データ入力
S720
項目指定
S730
文書データ→変換データ
S740
分離記号挿入
S750
変換データ記憶
S760
文書ベクトル生成
S770
文書ベクトル記憶
S780
文書分類
S790
分類結果記憶
END

【書類名】　　要約書

【要約】

【課題】　文書間の類似性に基づいて文書分類をおこなう際、操作者の意図を反映する文書分類をおこなうことを課題とする。

【解決手段】　一つまたは複数の項目から構成された文書データを入力する入力部４０１と、入力された文書データを構成する前記項目を指定する指定部４０２と、指定された項目に対応するデータのみの内容となるように前記文書データを変換する変換部４０３と、変換された変換データをもちいて文書を分類する分類部４０５とを備える。

【選択図】　　図４

【書類名】　職権訂正データ
【訂正書類】　特許願

＜認定情報・付加情報＞
【特許出願人】　申請人
【識別番号】　000006747
【住所又は居所】　東京都大田区中馬込１丁目３番６号
【氏名又は名称】　株式会社リコー

出　願　人　履　歴　情　報

識別番号　　〔０００００６７４７〕

１．変更年月日　　１９９０年　８月２４日
［変更理由］　新規登録
住　所　東京都大田区中馬込１丁目３番６号
氏　名　株式会社リコー